iDASH Healthcare Privacy Protection Challenge

Fei Yu
feiy@stat.cmu.edu

Carnegie Mellon University

24 March 2014

Overview

Task

Select the K most significant SNPs differentially-privately.

* Setting: case-control study.
* Input data: genotype data (e.g., AA, AT, TT) for cases, minor allele frequencies for controls.
* Ranking significance: p-value corresponding to Pearson's χ^2 test of association between SNP and phenotype.
* Performance evaluation: the proportion of significant SNPs recovered.

Overview

* Method is based on the exponential mechanism.
* Two variations of the method. Pros and cons.

Definitions

Differential privacy

Let $\mathcal{D}$ denote the set of all data sets. Write $D \sim D'$ if D and D' differ in one individual. A randomized mechanism $\mathcal{K}$ is ϵ-differentially private if, for all $D \sim D'$ and for any measurable set $S \subset \mathbb{R}$,

$$\frac{Pr(\mathcal{K}(D) \in S)}{Pr(\mathcal{K}(D') \in S)} \leq e^{\epsilon}.$$

Sensitivity

The sensitivity of a function $f : \mathcal{D}^N \to \mathbb{R}^d$, where $\mathcal{D}^N$ denotes the set of all databases with N individuals, is the smallest number $S(f)$ such that

$$||f(D) - f(D')||_1 \leq S(f),$$

for all data sets $D, D' \in \mathcal{D}^N$ such that $D \sim D'$.

Exponential mechanism

McSherry and Talwar (2007): Given $D = \{\text{SNP}_i\}_{i=1}^{M}$, ε_q^ϵ is a r.v. with

$$\begin{aligned}\Pr(\varepsilon_q^\epsilon(D) = i) &\propto \exp\left(\frac{\epsilon q(D,i)}{2\Delta_q}\right)\mu(i)\\ &\propto \exp\left(\frac{\epsilon q(D,i)}{2s}\right)\end{aligned}$$

where

$$\begin{aligned}q(D,i) &= \text{ the score for SNP}_i\\ s &= \text{ the sensitivity of } q(D,\cdot)\\ \mu(i) &= 1/M.\end{aligned}$$

ε_q^ϵ is ϵ-differentially private.

Exponential mechanism

We can use any scoring function $q(D, \cdot)$ with the exponential mechanism. Examples:

1. χ^2 statistic
2. Hamming distance (Johnson and Shmatikov 2013)

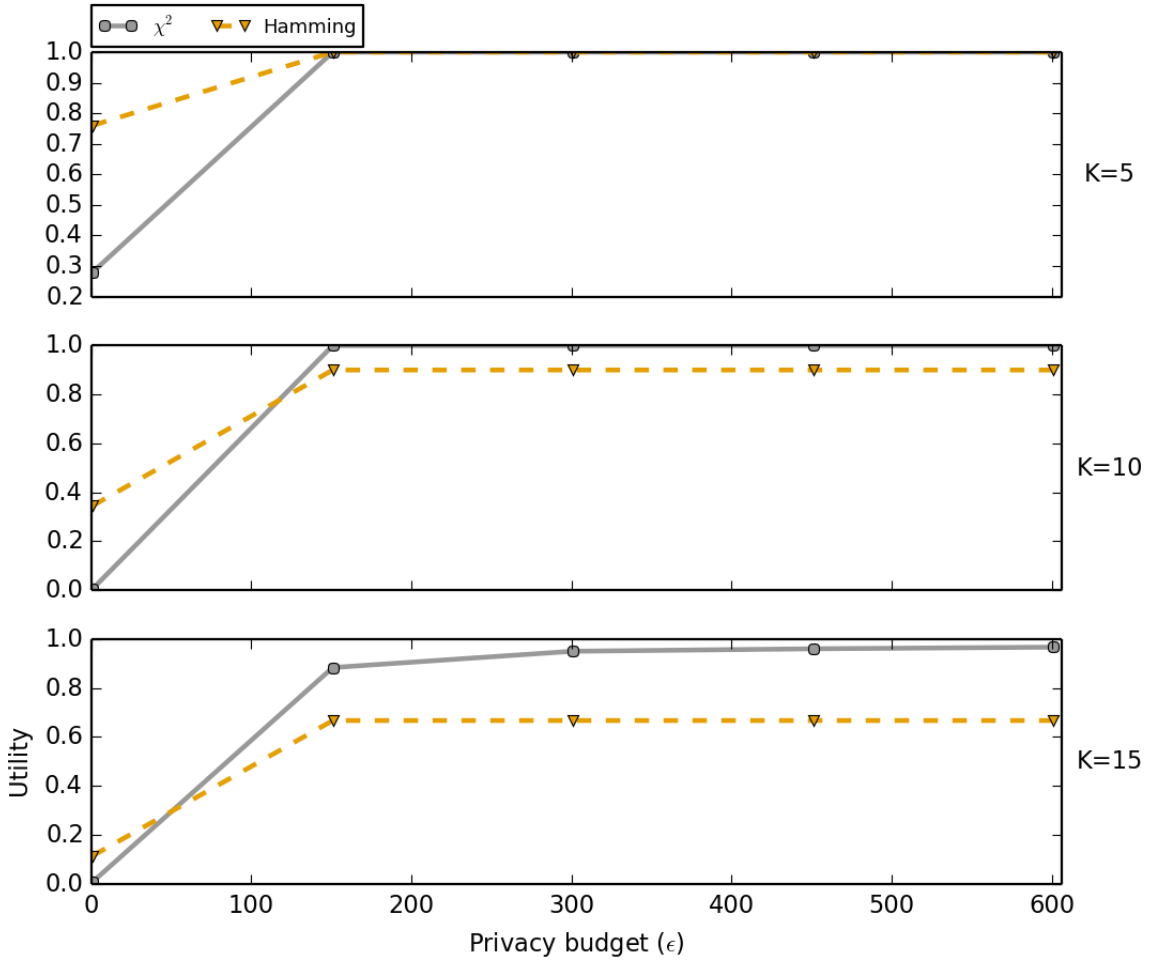
Extending the exponential mechanism

Johnson and Shmatikov (2013): selecting the K most significant SNPs (`LocSig`).

1. Initialize $\mathcal{S} = \emptyset$ and $q_i =$ score of SNP$_i$.
2. Set $w_i = \exp\left(\frac{\epsilon q_i}{2Ks}\right)$ and $\Pr(\varepsilon_q^\epsilon(D) = i) = w_i \Big/ \sum_{j=1}^{M} w_j$.
3. Sample $j \sim \varepsilon_q^\epsilon(D)$. Add SNP$_j$ to $\mathcal{S}$. Set $q_j = -\infty$.
4. If $|\mathcal{S}| < K$, return to Step 2. Otherwise, output $\mathcal{S}$.

`LocSig` is ϵ-differentially private (Yu et al. 2014).

Performance of different scoring functions



* Hamming (distance) outperforms χ^2 when ϵ is small.
* Utility of Hamming may plateur before it reaches 1.0. (Why?)

Setup

Assumptions:

* # of cases = # of controls = $N/2$.
* Case data are private but control data are known.

Setup

Summarizing a SNP:

* Genotype table is not available. We only know the genotypes of the cases:

Genotype	0	1	2	
Case	g_0	g_1	g_2	$N/2$

* Derived allelic table:

Allele	0	1	
Case	n_{00}	n_{01}	N
Control	n_{10}	n_{11}	N
	n_0	n_1	$2N$

Using χ^2 statistic as score

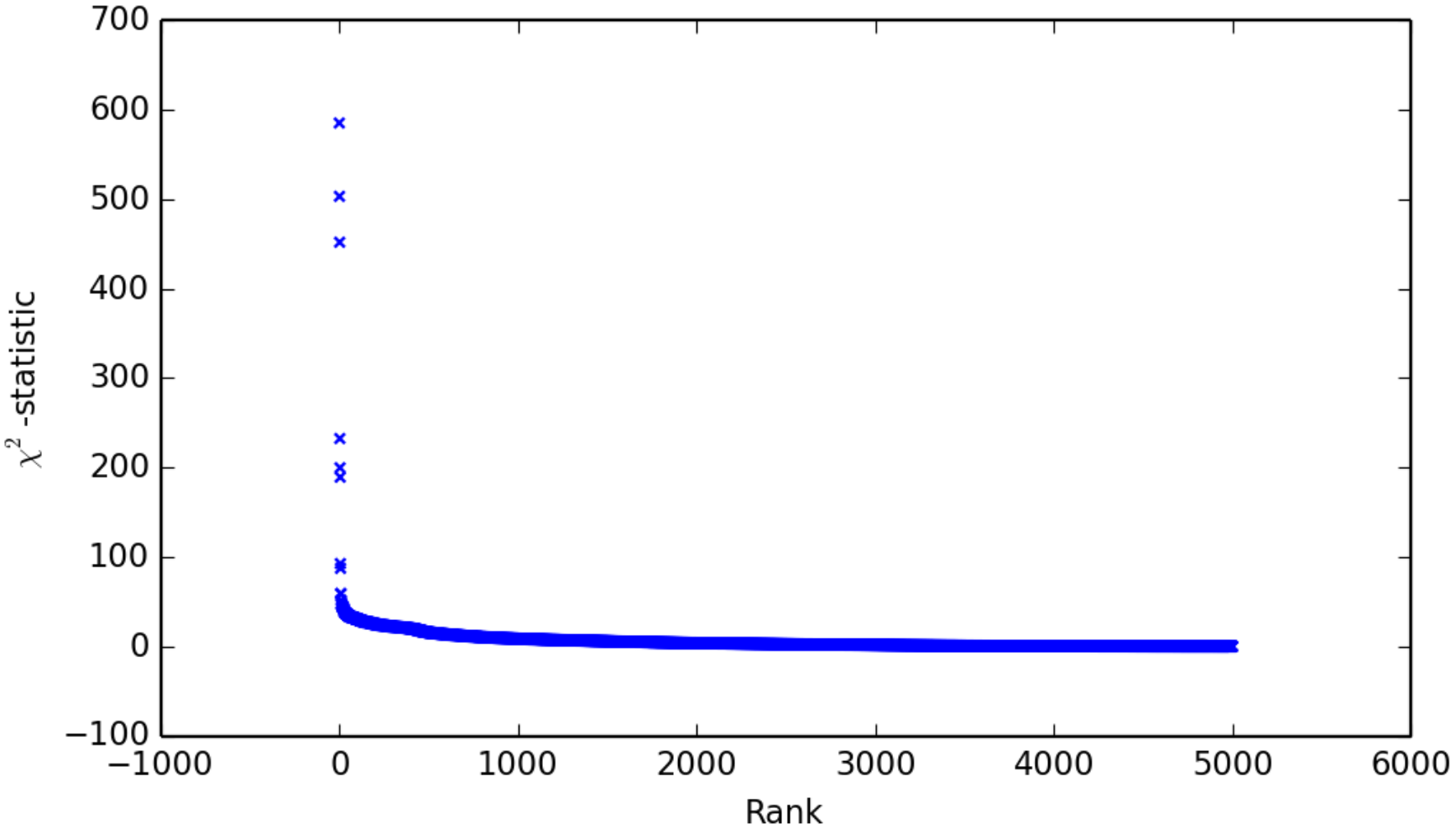
* Pearson's χ^2 statistics are used to rank significance of SNPs.
* Higher utility is attainable by increasing ϵ.

* Sensitivity of the Pearson's χ^2 statistic of an allelic table with positive margins, $N/2$ cases and $N/2$ controls is

$$\frac{8N^2}{(N+3)(N+1)}\left(1-\frac{2}{N}\right) \quad \text{when } N \geq 3.$$

See Yu et al. (2014).

χ^2 statistic vs. ranking



Using χ^2 statistic as score

* Pearson's χ^2 statistics are used to rank significance of SNPs.
* Higher utility is attainable by increasing ϵ.

* Sensitivity of the Pearson's χ^2 statistic of an allelic table with positive margins, $N/2$ cases and $N/2$ controls is

$$\frac{8N^2}{(N+3)(N+1)}\left(1-\frac{2}{N}\right) \quad \text{when } N \geq 3.$$

See Yu et al. (2014).

Using Hamming distance as score

$$
\begin{array}{ccccc}
D \sim & D_1 & \sim \cdots \sim & D_{n-1} \sim & D_n \\
\Downarrow & \Downarrow & & \Downarrow & \Downarrow \\
p & p_1 & \cdots & p_{n-1} & p_n \\
\text{(sig)} & \text{(sig)} & & \text{(sig)} & \text{(not sig)}
\end{array}
$$

* Score > 0 only when $D \in \mathcal{D}$ is significant.
* SNP significance ordering resulting from Hamming distance could be different than that resulting from χ^2 statistic.
* Sensitive to the choice of the threshold p-value.
* No genotype data for controls: necessary to assume controls are known.

Using Hamming distance as score

$$
\begin{array}{ccccc}
D \sim & D_1 & \sim \cdots \sim & D_{n-1} \sim & D_n \\
\Downarrow & \Downarrow & & \Downarrow & \Downarrow \\
p & p_1 & \cdots & p_{n-1} & p_n \\
\text{(sig)} & \text{(sig)} & & \text{(sig)} & \text{(not sig)}
\end{array}
$$

* Score > 0 only when $D \in \mathcal{D}$ is significant.
* SNP significance ordering resulting from Hamming distance could be <u>different</u> than that resulting from χ^2 statistic.
* Sensitive to the choice of the threshold p-value.
* No genotype data for controls: necessary to assume controls are known.

Using Hamming distance as score

$$
\begin{array}{ccccc}
D \sim & D_1 & \sim \cdots \sim & D_{n-1} \sim & D_n \\
\Downarrow & \Downarrow & & \Downarrow & \Downarrow \\
p & p_1 & \cdots & p_{n-1} & p_n \\
\text{(sig)} & \text{(sig)} & & \text{(sig)} & \text{(not sig)}
\end{array}
$$

* Score > 0 only when $D \in \mathcal{D}$ is significant.
* SNP significance ordering resulting from Hamming distance could be <u>different</u> than that resulting from χ^2 statistic.
* Sensitive to the choice of the threshold p-value.
* No genotype data for controls: necessary to assume controls are known.

Using Hamming distance as score

$$
\begin{array}{ccccc}
D \sim & D_1 & \sim \cdots \sim & D_{n-1} \sim & D_n \\
\Downarrow & \Downarrow & & \Downarrow & \Downarrow \\
p & p_1 & \cdots & p_{n-1} & p_n \\
\text{(sig)} & \text{(sig)} & & \text{(sig)} & \text{(not sig)}
\end{array}
$$

* Score > 0 only when $D \in \mathcal{D}$ is significant.
* SNP significance ordering resulting from Hamming distance could be <u>different</u> than that resulting from χ^2 statistic.
* Sensitive to the choice of the threshold p-value.
* No genotype data for controls: necessary to assume controls are known.

Finding the Hamming distance

$$
\begin{array}{ccccccc}
D \sim & D_1 & \sim \cdots \sim & D_{n-1} \sim & D_n \\
\Downarrow & \Downarrow & & \Downarrow & \Downarrow \\
p & p_1 & \cdots & p_{n-1} & p_n \\
\text{(sig)} & \text{(sig)} & & \text{(sig)} & \text{(not sig)}
\end{array}
$$

* Instead of examining all possible paths, follow the path of the greatest ascent or descent.
* The resulting path may not have the shortest Hamming distance.

Finding the Hamming distance

Partial genotype table

Genotype	0	1	2	
Case	g_0	g_1	g_2	$N/2$

Derived allelic table

Allele	0	1	
Case	n_{00}	n_{01}	N
Control	n_{10}	n_{11}	N
	n_0	n_1	$2N$

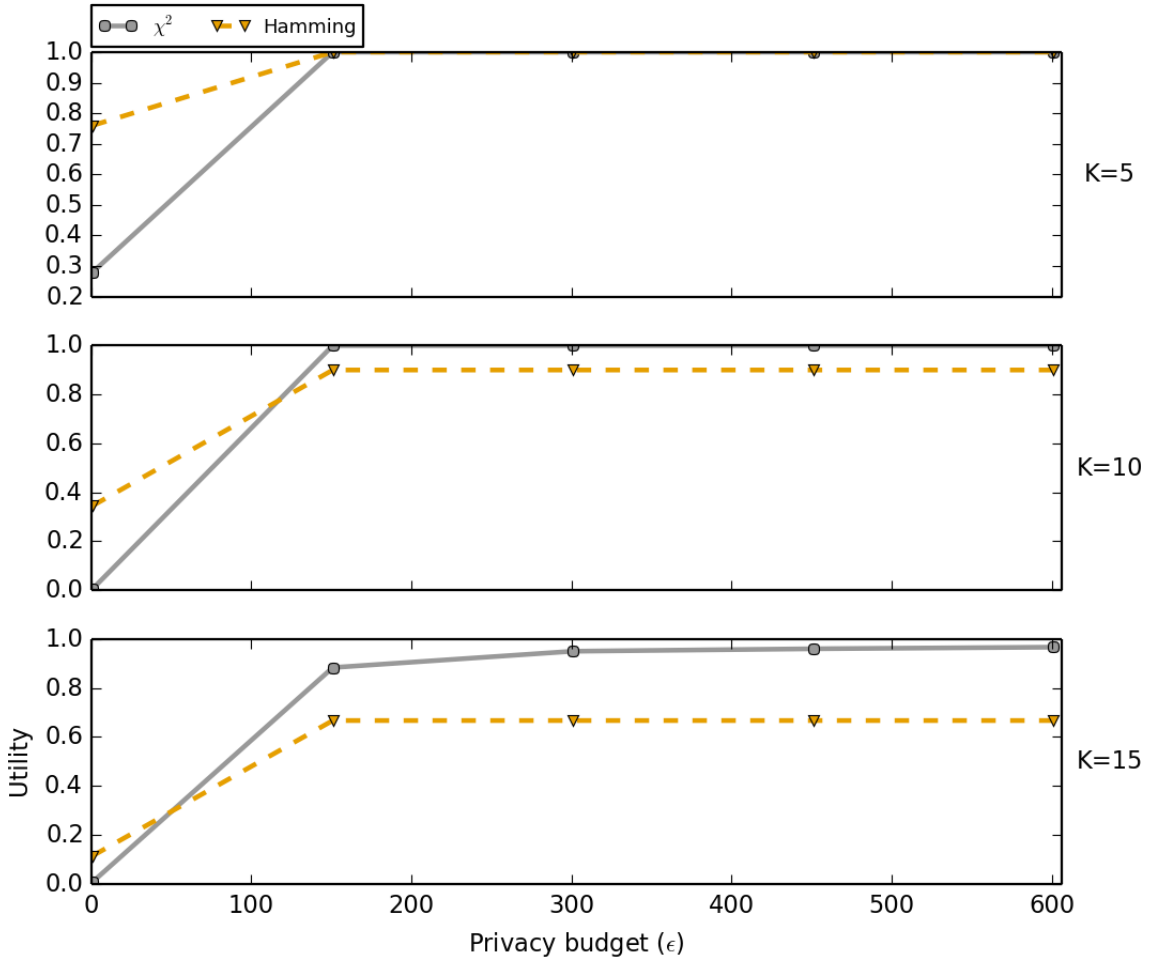
$$\chi^2 = \frac{2N(n_{00} - n_{10})^2}{n_0 n_1} = \frac{2N(2g_0 + g_1 - n_{10})^2}{(2g_0 + g_1 + n_{10})(N - 2g_0 - g_1 - n_{10})}$$

$$\nabla\chi^2 = \left(\frac{\partial}{\partial g_0}\chi^2, \frac{\partial}{\partial g_1}\chi^2\right)$$

$$\frac{\partial}{\partial g_0}\chi^2 = 2\frac{\partial}{\partial g_1}\chi^2$$

$$\frac{\partial}{\partial g_1}\chi^2 \propto \left(\frac{n_{00}}{n_0}\frac{n_{11}}{n_1} - \frac{n_{01}}{n_1}\frac{n_{10}}{n_0}\right)\left(\frac{n_{10}}{n_0} + \frac{n_{01}}{n_1}\right)$$

Performance of different scoring functions



* Hamming (distance) outperforms χ^2 when ϵ is small.
* Utility of Hamming may plateur before it reaches 1.0. (Why?)

Comparison of scoring functions

	χ^2	Hamming
Computation	Trivial	Expensive
Sensitivity	Nontrivial; may use upper bounds	1
Stable	Yes	Not always

Summary

* Extending exponential mechanism — `LocSig`
* χ^2 statistic as score
* Hamming distance as score
* Compare different scoring functions

References

- Johnson, Aaron, and Vitaly Shmatikov. 2013. "Privacy-preserving data exploration in genome-wide association studies". In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1079–1087.
- McSherry, Frank, and Kunal Talwar. 2007. "Mechanism Design via Differential Privacy". *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* (): 94–103.
- Yu, Fei, et al. 2014. "Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies". *Journal of Biomedical Informatics* (). arXiv: 1401.5193.